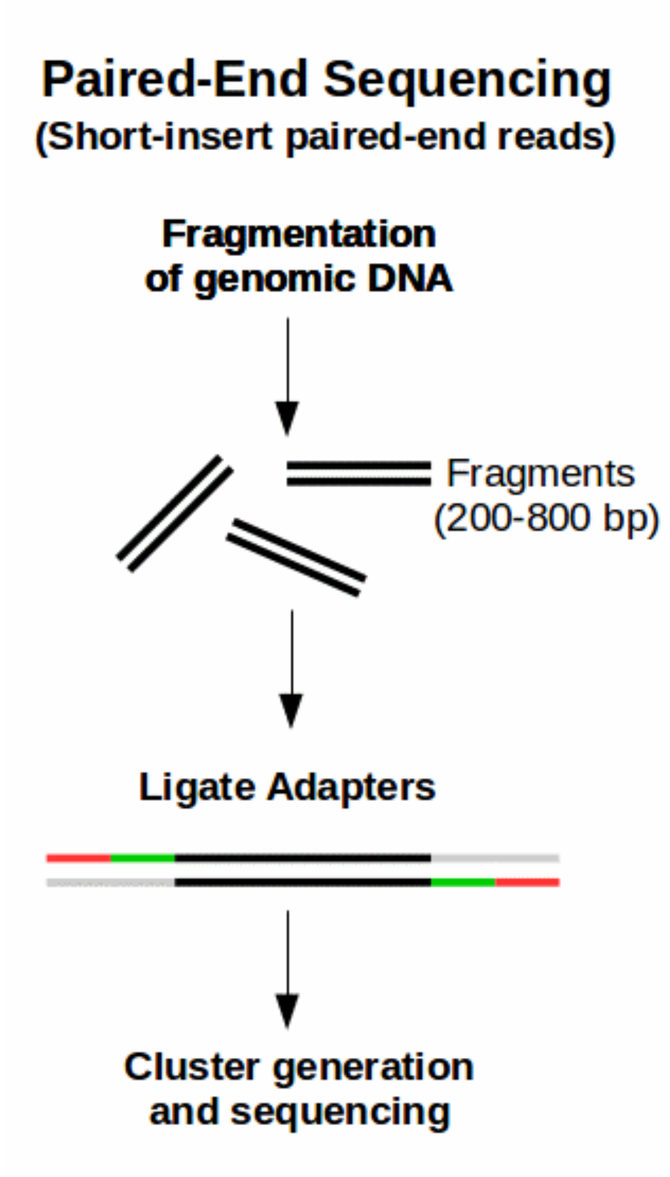
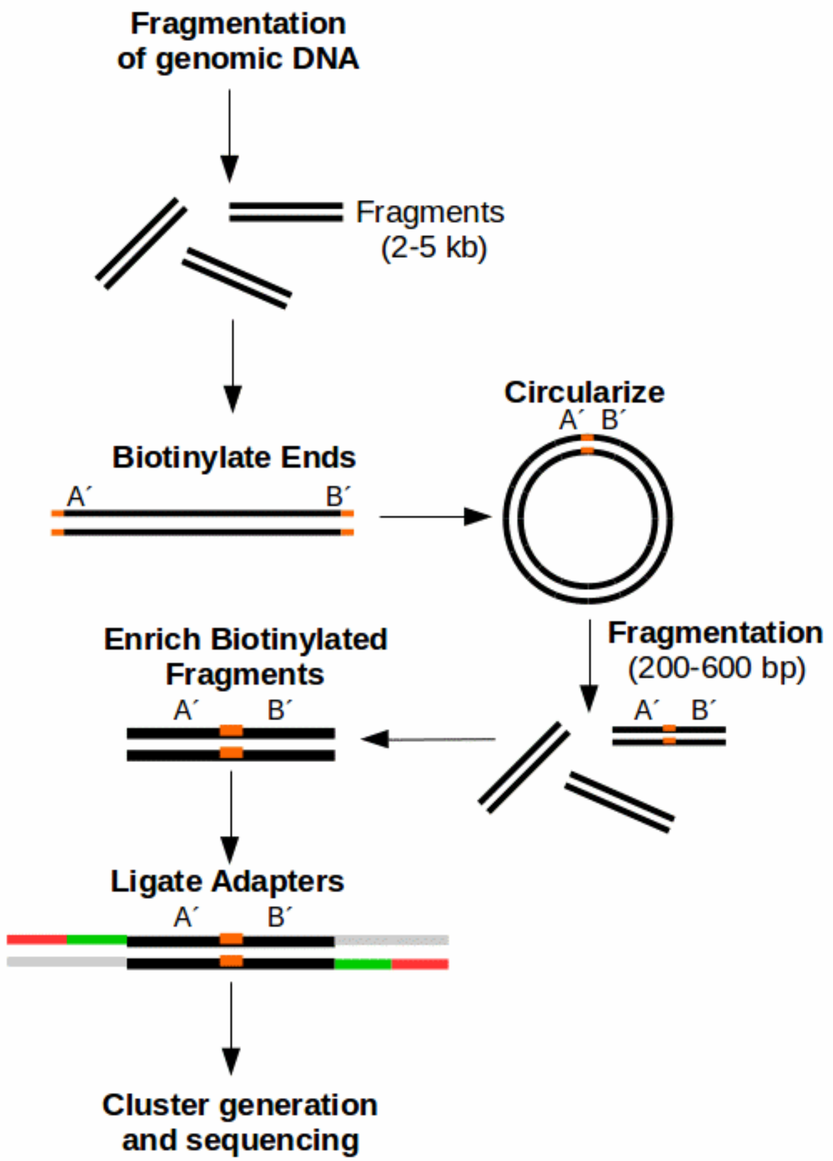
Three invariant Hi-C interaction patterns: Applications to genome assembly

Introduction


Paired-End Sequencing
(Short-insert paired-end reads)
Fragmentation
of genomic DNA
Fragments
(200-800 bp)
Ligate Adapters
Cluster generation
and sequencing

Mate Pair Sequencing



Also useful

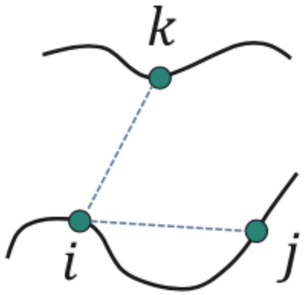
- (1) Haplotype phasing
  - Physical separation of homologous chromosomes in the nucleus
  - -> Better SNP linkage information over large distances
- (2) Metagenome deconvolution
  - Connect contigs because of long-distance linkage information
  - Genomic interactions between cells is extremely low)
- (3) Cancer genomes
  - Large structural variations are especially difficult to measure with short reads (e.g. rearrangements)
  - In Hi-C, structural variations can be detected since they appear to deviate from standard Hi-C patterns [32,33]

Invariant Hi-C patterns

- Since the 3D organization of a genome reflects its functional state, it is not surprising that 3D genome organization differs between species.
    - In fact, 3D genome organization varies between cell-types [8,20],
    - Along different stages of the cell-cycle [44–46],
    - And even within homogenous populations of synchronized cells [46]
- Despite this, certain aspects of 3D genome organization, as measured by Hi-C, are universal [23, 24] (refer as ***invariant patterns***)
- In fact, these patterns are so robust and ubiquitous that they are used to evaluate the quality of Hi-C experiments and check for experimental artifacts [34]
- We will focus on level of individual loci

Invariant pattern I: Intrachromosomal interaction enrichment

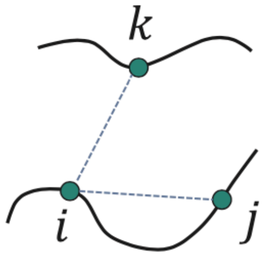
- Chromosome territories, random positioning of chromosomes (not single-cell)
- Typically the ratio between intrachromosomal (cis) and interchromosomal (trans) is used as a quality metric for Hi-C
- The underlying logic is that general random noise will affect the interaction matrix uniformly
- $chr(i) = chr(j), chr(i) \neq chr(k) \Rightarrow p_{int}(i,j) > p_{int}(i,k)$

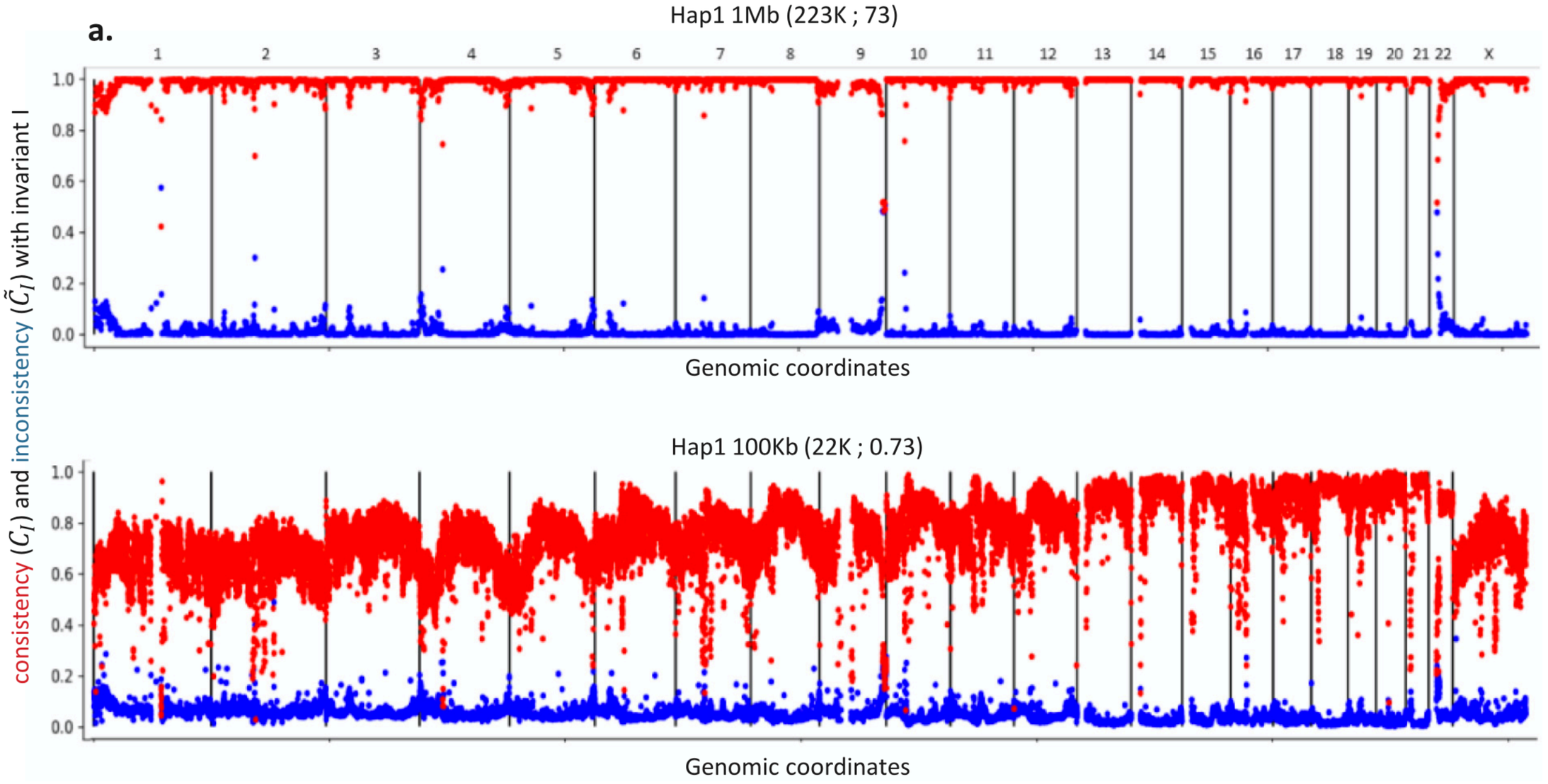


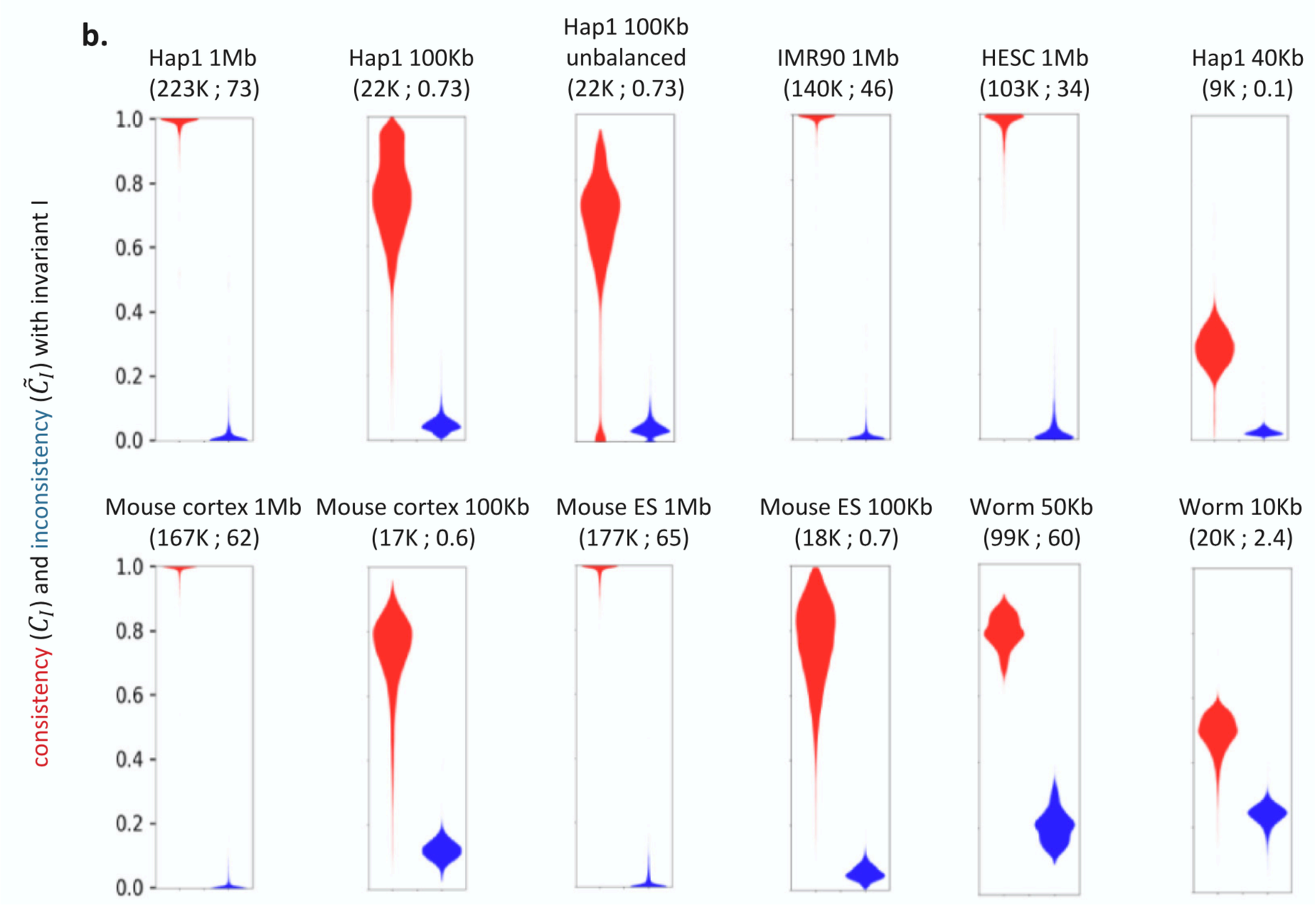
- In other words, if j is on the same chromosome as i but k is not, i will interact more frequently with j than with k

Consistent: $chr(i) = chr(j), chr(i) \neq chr(k) \Rightarrow p_{int}(i,j) > p_{int}(i,k)$

Inconsistent: $chr(i) = chr(j), chr(i) \neq chr(k) \Rightarrow p_{int}(i,j) < p_{int}(i,k)$

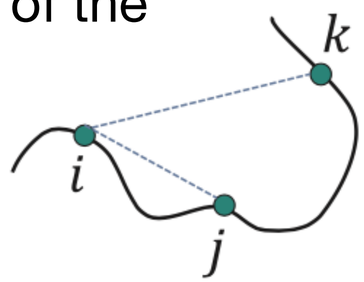





b.
Hap1 1Mb
(223K ; 73)
Hap1 100Kb
(22K ; 0.73)
Hap1 100Kb
unbalanced
(22K ; 0.73)
IMR90 1Mb
(140K ; 46)
HESC 1Mb
(103K ; 34)
Hap1 40Kb
(9K ; 0.1)
Mouse cortex 1Mb
(167K ; 62)
Mouse cortex 100Kb
(17K ; 0.6)
Mouse ES 1Mb
(177K ; 65)
Mouse ES 100Kb
(18K ; 0.7)
Worm 50Kb
(99K ; 60)
Worm 10Kb
(20K ; 2.4)
1.0
0.8
0.6
0.4
0.2
0.0
consistency (C_I) and inconsistency ($\tilde{C}_I$) with invariant I

Invariant pattern II: distance-dependent interaction decay

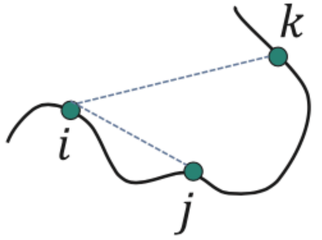
- Locus interacts more frequently with loci which are nearby in the genomic sequence than with far away loci
- Inherent feature of many polymer physics models (will interact frequently by random)
- Offtop: exact details of the distance dependence can be used to suggest properties of the underlying polymer

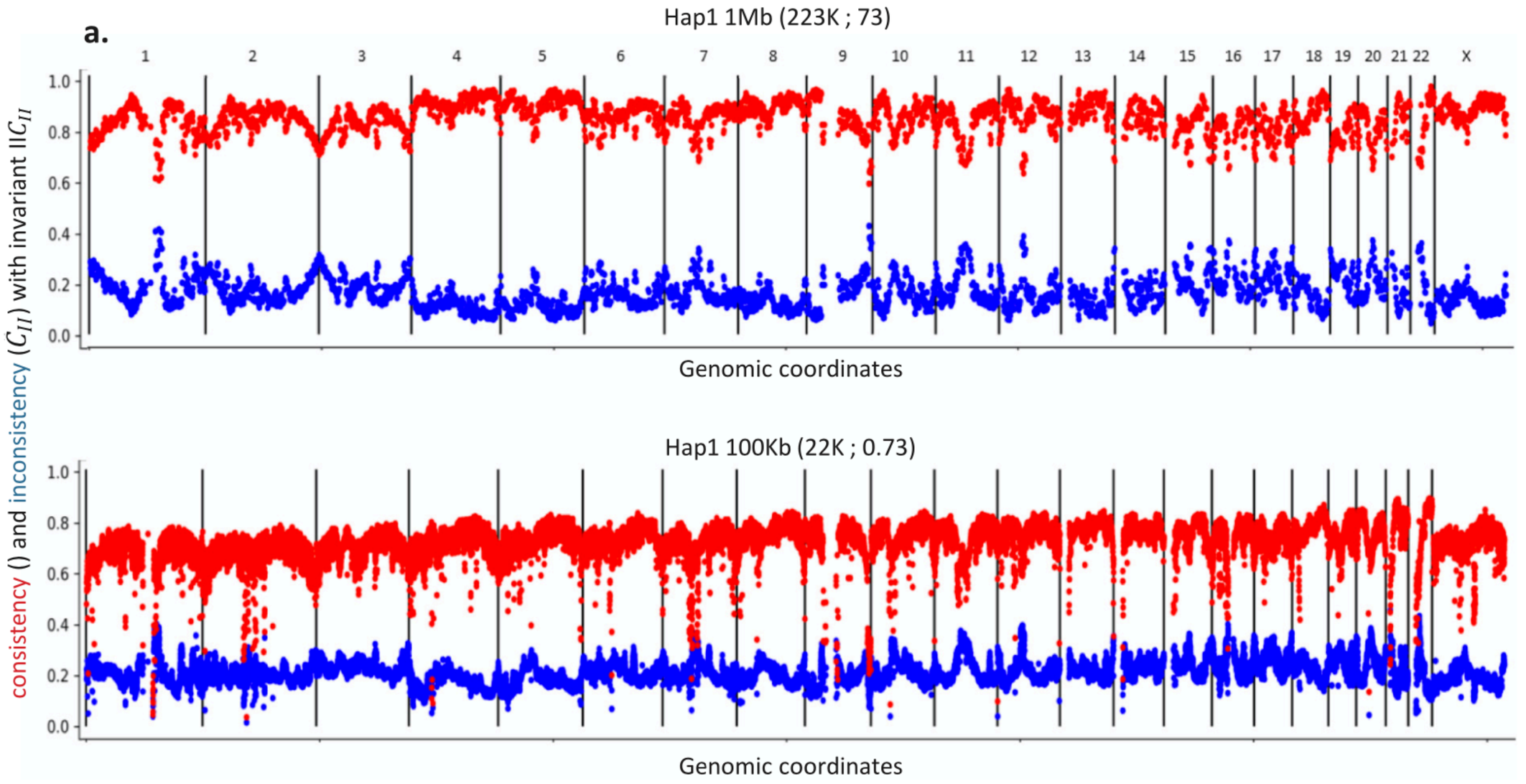


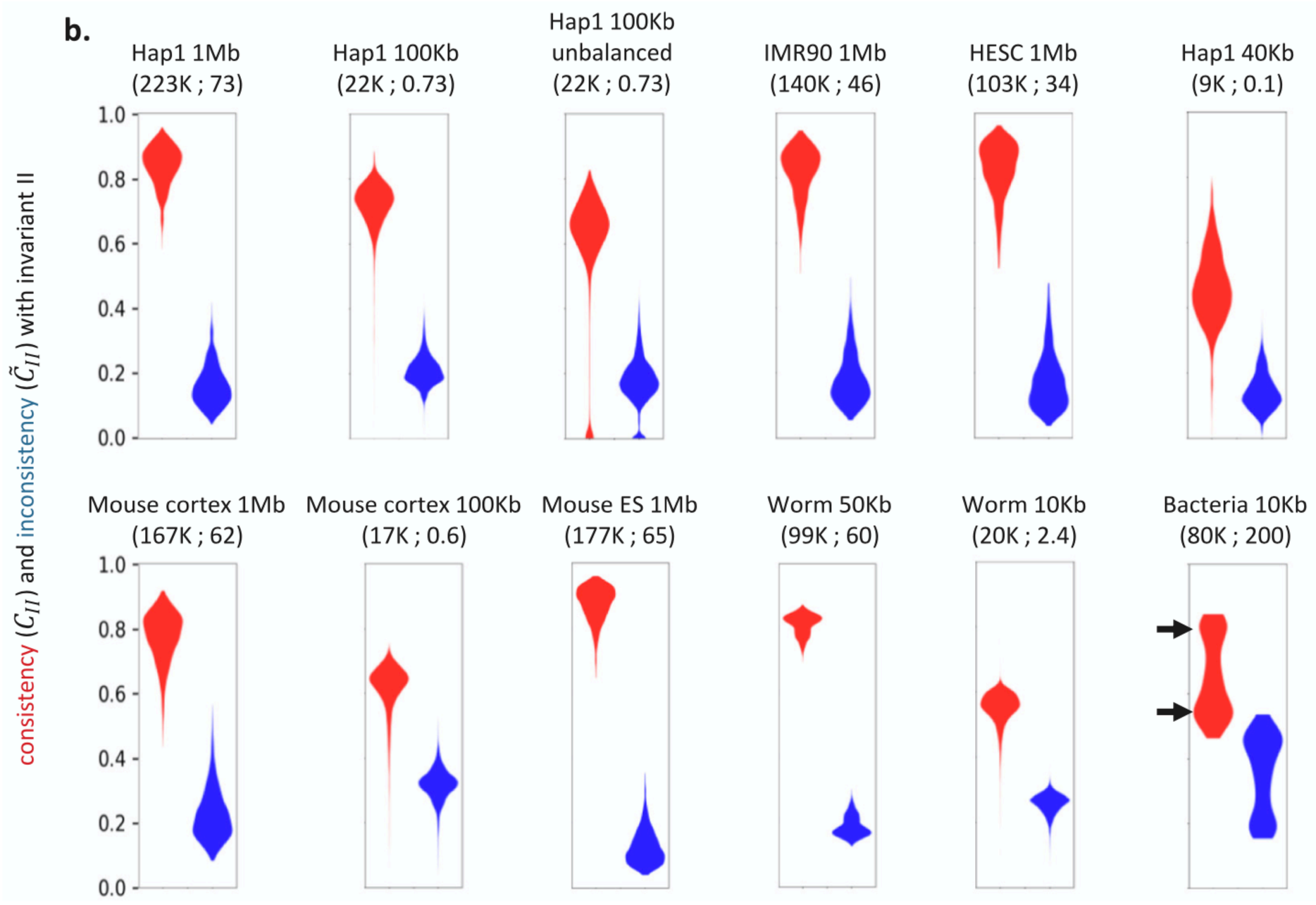
- $|i-j| < |i-k| \Rightarrow p_{int}(i,j) > p_{int}(i,k)$
- In other words, if j is closer (in genomic distance) to i than k is, i will interact more frequently with j
- Offtop: invariant II implies invariant I if we define that being on a different chromosome is equivalent to being infinitely far in genomic sequence

Consistent: $|i-j| < |i-k| \Rightarrow p_{int}(i,j) > p_{int}(i,k)$

Inconsistent: $|i-j| < |i-k| \Rightarrow p_{int}(i,j) < p_{int}(i,k)$

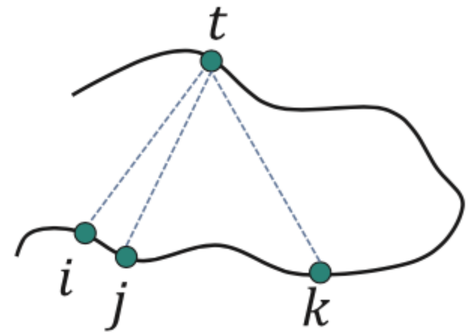





b.
Hap1 1Mb
(223K ; 73)
Hap1 100Kb
(22K ; 0.73)
Hap1 100Kb
unbalanced
(22K ; 0.73)
IMR90 1Mb
(140K ; 46)
HESC 1Mb
(103K ; 34)
Hap1 40Kb
(9K ; 0.1)
Mouse cortex 1Mb
(167K ; 62)
Mouse cortex 100Kb
(17K ; 0.6)
Mouse ES 1Mb
(177K ; 65)
Worm 50Kb
(99K ; 60)
Worm 10Kb
(20K ; 2.4)
Bacteria 10Kb
(80K ; 200)
1.0
0.8
0.6
0.4
0.2
0.0
consistency (C_{II}) and inconsistency ($\tilde{C}_{II}$) with invariant II

Invariant pattern III: local interaction smoothness

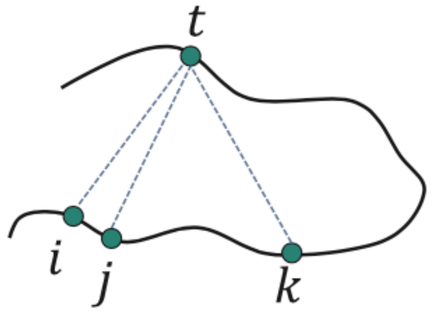
- Tendency of nearby loci to share similar interactions
- The physical basis of this invariant is intuitive: given the set of interactions of a genomic locus A, if we consider a locus B that is sufficiently close to A it will also be close to the neighbors of A
- This invariant is informally used to assess experimental artifacts in Hi-C experiments, and one of the criteria for successful removal of biases (e.g. by matrix balancing) is the visual smoothness of the resulting matrix [49, 50]
- This invariant pattern has not been stated explicitly, to the best of our knowledge
- $|i-j| < |i-k| \Rightarrow |p_{int}(i,t) - p_{int}(j,t)| < |p_{int}(i,t) - p_{int}(k,t)|$



- In other words, if j is closer to i than k is, we expect the interaction probability of (j,t) to be closer to that of (i,t) than the interaction probability of (k,t) is to that of (i,t)
- We considered positions j = i + 1, k = i + 10 (units are bins) as well as every same-chromosome locus t

Consistent: $|i-j| < |i-k| \Rightarrow |p_{int}(i,t) - p_{int}(j,t)| < |p_{int}(i,t) - p_{int}(k,t)|$

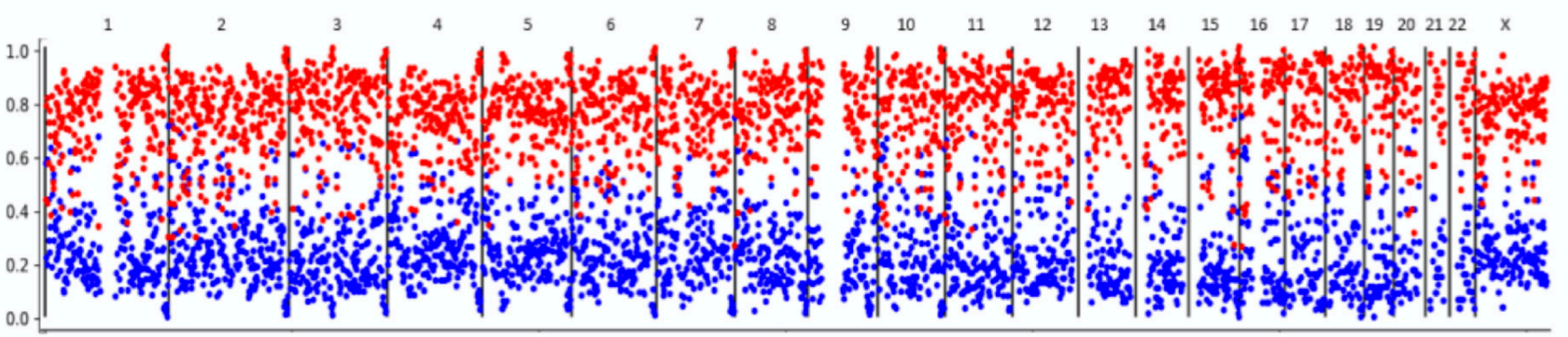
Inconsistent: $|i-j| < |i-k| \Rightarrow |p_{int}(i,t) - p_{int}(j,t)| > |p_{int}(i,t) - p_{int}(k,t)|$



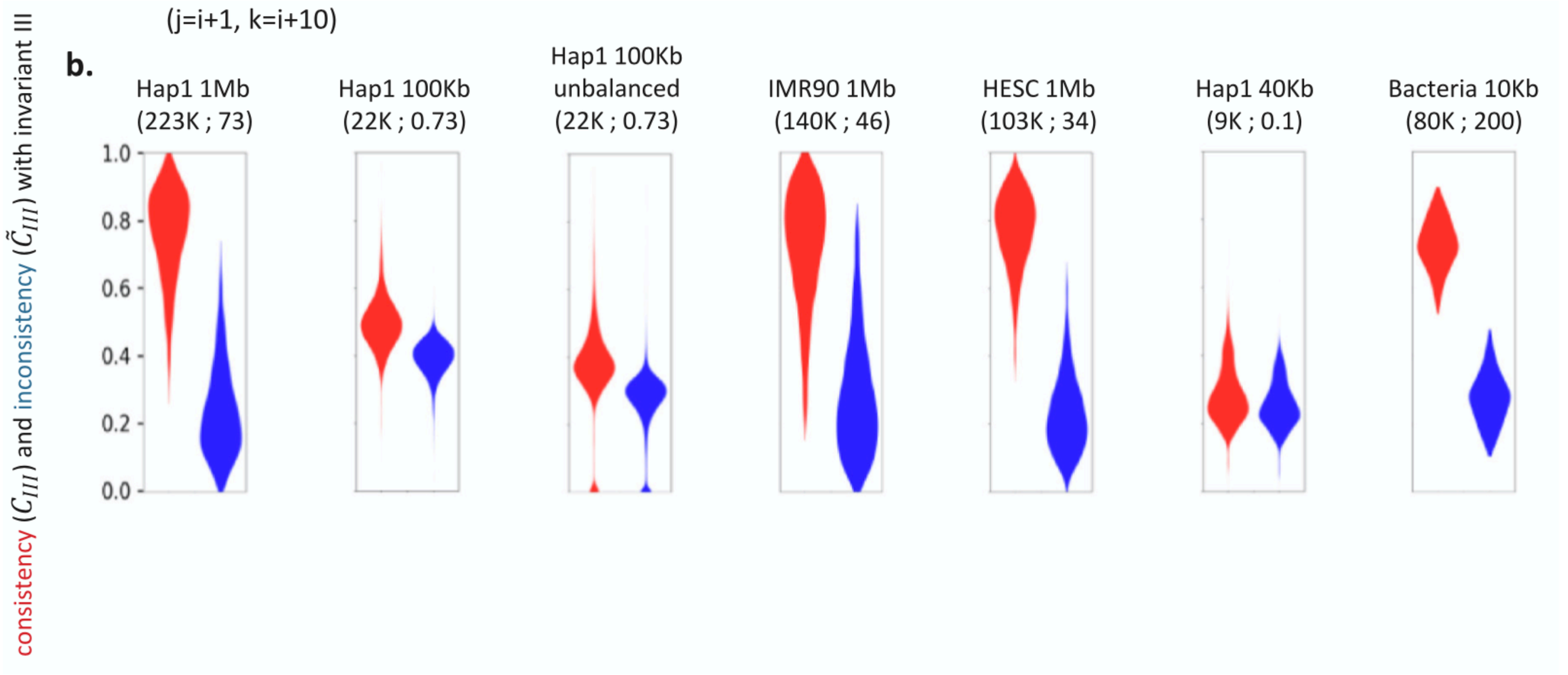
a.

(j=i+1, k=i+10)

Hap1 1Mb (223K ; 73)



consistency (C_{III}) and inconsistency ($\tilde{C}_{III}$) with invariant III


(j=i+1, k=i+10)
b.
consistency (C_{III}) and inconsistency ($\tilde{C}_{III}$) with invariant III
Hap1 1Mb
(223K ; 73)
Hap1 100Kb
(22K ; 0.73)
Hap1 100Kb
unbalanced
(22K ; 0.73)
IMR90 1Mb
(140K ; 46)
HESC 1Mb
(103K ; 34)
Hap1 40Kb
(9K ; 0.1)
Bacteria 10Kb
(80K ; 200)
1.0
0.8
0.6
0.4
0.2
0.0

Smoothness

- $$s(i) = \frac{1}{|T|} \sum_{t} |p_{int}(i,t) - p_{int}(k,t)| - |p_{int}(i,t) - p_{int}(j,t)|$$

- Remarkably, we find that s values are overwhelmingly positive (0.999 of the bins are positive), in contrast to C_{III} values

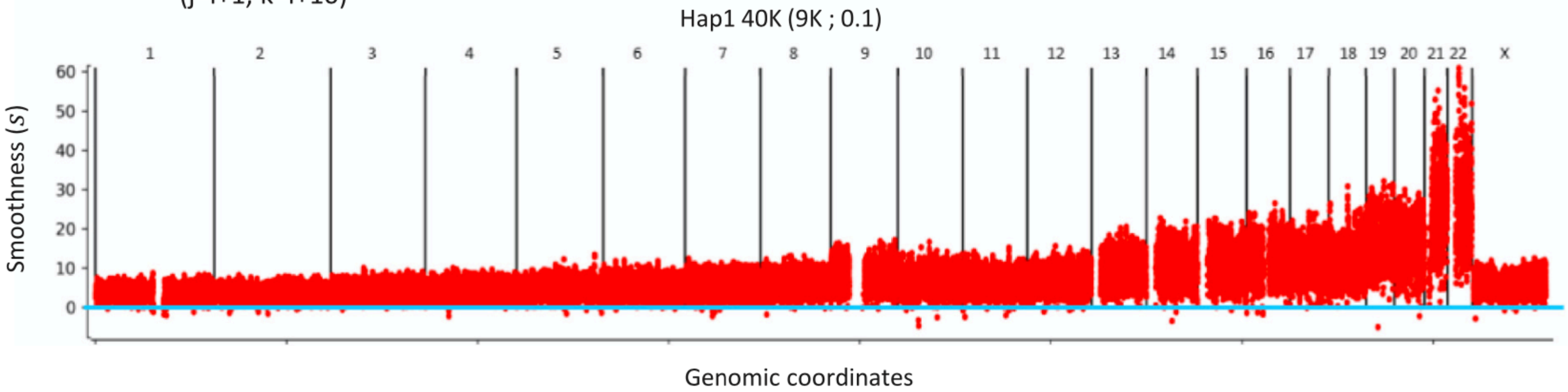
- We suggest this difference is due to many inconsistent loci having low interaction probabilities and thus a lower weight than consistent loci

- We suggest using s **for practical applications** such as detecting assembly errors and structural variation

c.

$$s(i) = \frac{1}{|T|}\sum_{t} |p_{int}(i,t) - p_{int}(k,t)| - |p_{int}(i,t) - p_{int}(j,t)|$$

(j=i+1, k=i+10)



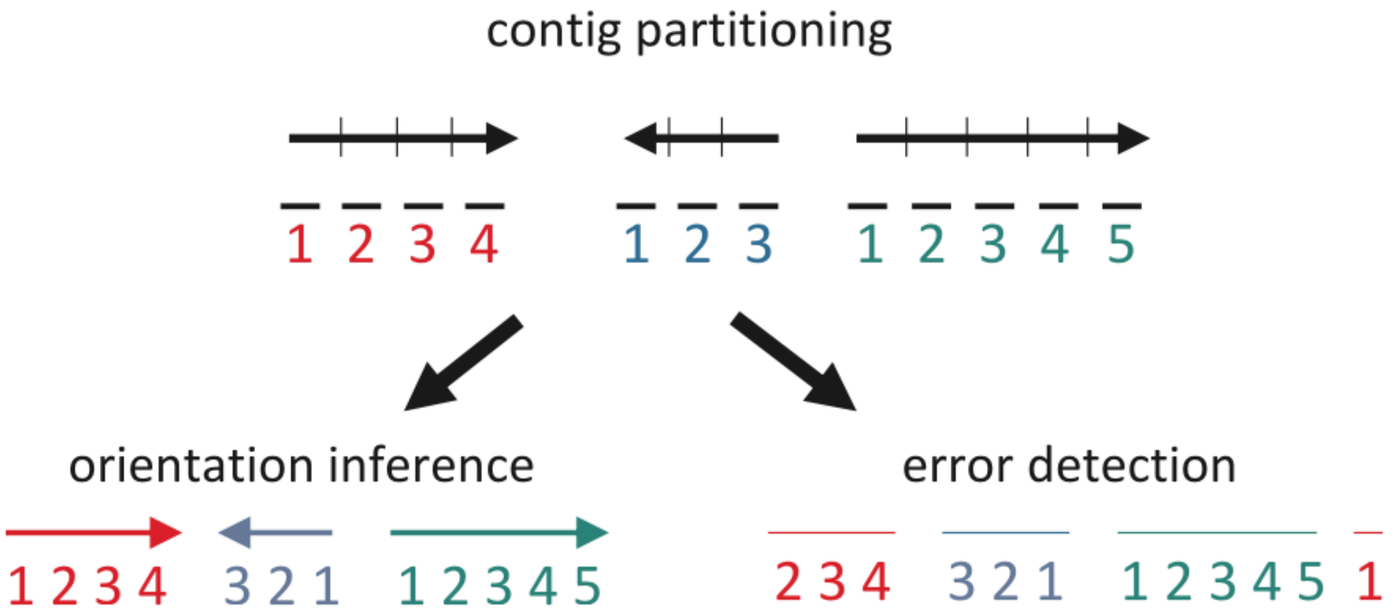
General workflow

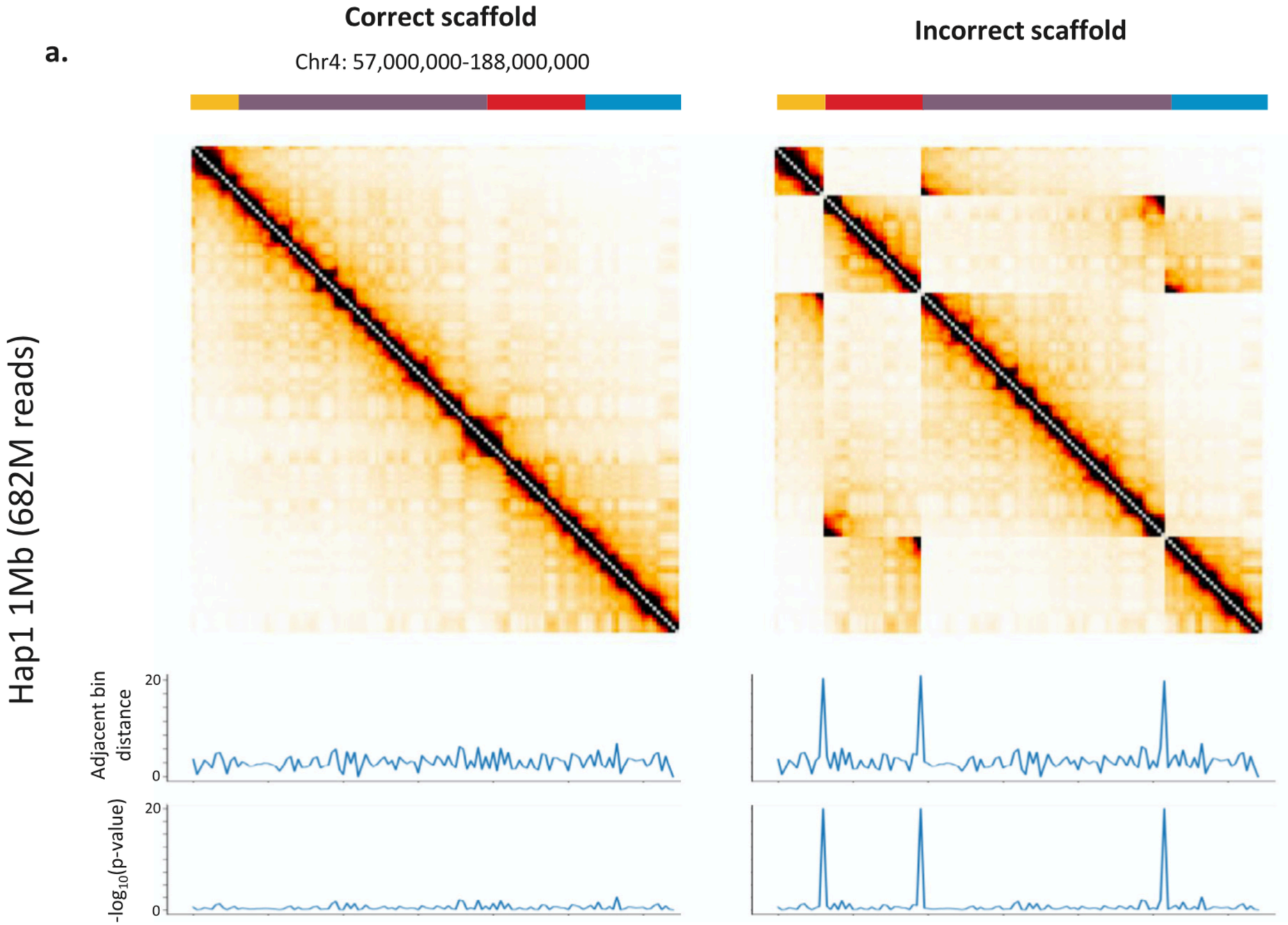
- We recommend the following general workflow:

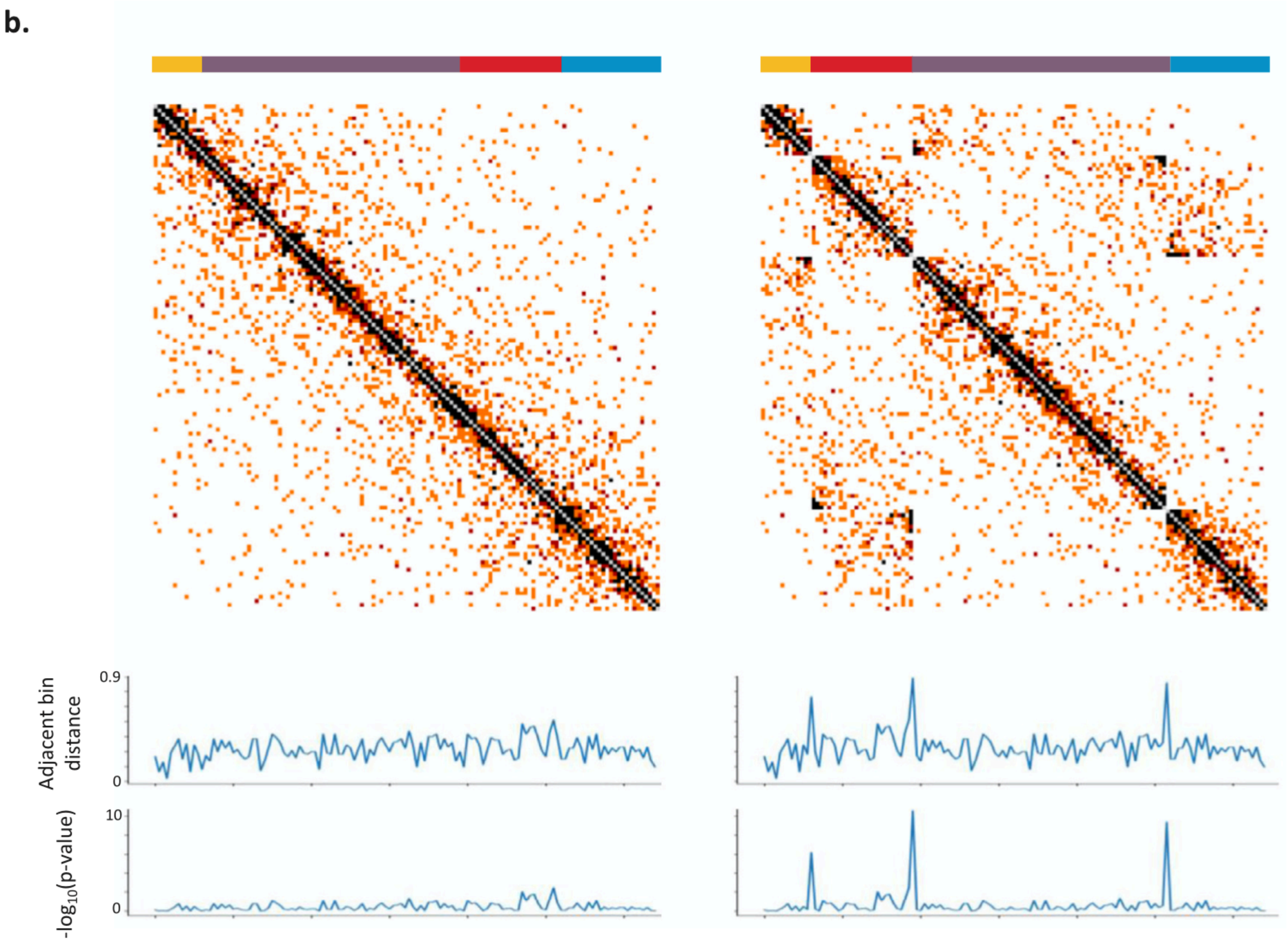
    1. assemble initial set of contigs;
    2. map Hi-C data to contigs;
    3. partition contigs into bins (sub-contigs) of a set size, allowing for identification of problematic contigs, intrinsic evaluation of scaffolding, and contig orientation;
    4. remove problematic bins and perform Hi-C correction using any available correction method (e.g. matrix balancing); if chromosome number is unknown, estimate chromosome number using DNA Triangulation bootstrapped clustering method;
    5. after chromosome number is chosen, partition contigs into chromosomes using clustering;
    6. identify and remove problematic contigs based on clustering results;
    7. scaffold each chromosome separately using probabilistic model;
    8. remove problematic contigs;
    9. evaluate results using orthogonal data and the invariant patterns.

- We next provide details on key points of this workflow

DNA triangulation


contig partitioning
1 2 3 4
1 2 3
1 2 3 4 5
orientation inference
1 2 3 4
3 2 1
1 2 3 4 5
error detection
2 3 4
3 2 1
1 2 3 4 5
1


a.
Correct scaffold
Chr4: 57,000,000-188,000,000
Incorrect scaffold
Hap1 1Mb (682M reads)
Adjacent bin distance
20
0
-log10(p-value)
20
0


b.
Hap1 1Mb (0.1M reads)
Adjacent bin distance
0.9
0
-log10(p-value)
10
0